Analyse de séquences

Partie III : Alignement multiple

Sèverine Bérard

AMAP - Université Montpellier 2

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
- Éditeurs de MSA
- Références

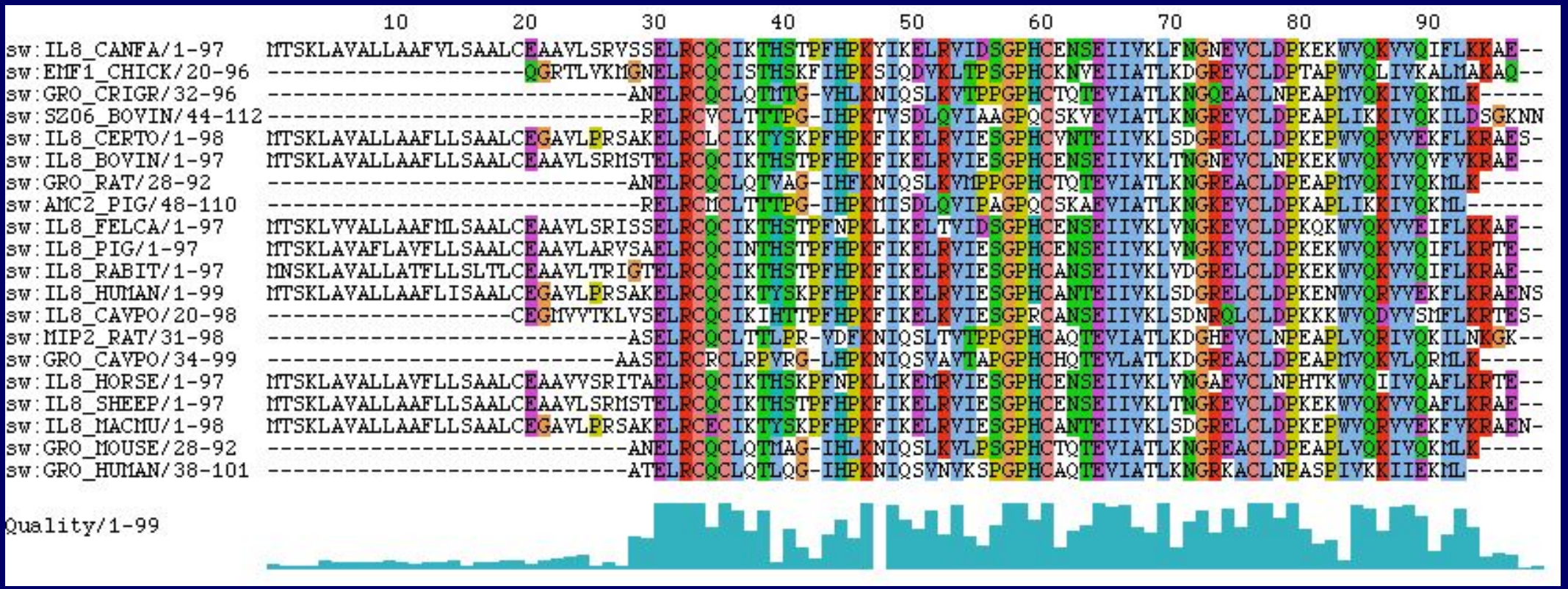
Alignement multiple

- Permet d'aligner plusieurs séquences simultanément
- En général pour les protéines
- Alignements faits à la main par les experts biologistes
- Généralisation naturelle de l'alignement 2 à 2 mais beaucoup plus complexe en terme de calcul

Problème : Étant données k séquences $s_1, s_2, \ldots, s_k$, trouver le meilleur alignement multiple pour ces séquences

- Ce pb est NP-complet ($\Rightarrow \nexists$ de solution exacte en tps raisonnable)

Exemples d'alignement



Intérêt des alignements multiples

- Recherche dans les banques $\Rightarrow$ plrs séquences similaires à la requête
  Il est naturel de vouloir aligner ces séquences entre elles

- MSA détectent les régions qui ont été conservées lors de l'évolution
  Très svt des domaines associés à une fonction clé de la molécule

- Plusieurs protéines de fonctions similaires dans différentes espèces
  $\rightarrow$ Quelles parties semblables ? $\Rightarrow$ Consensus/Profil
  $\rightarrow$ Quelles parties différentes ?

- Permet de trouver d'autres membres d'une famille de protéines

- Séquençage de génomes (assemblage, recouvrement EST)

- Point de départ pour les analyses phylogénétiques

Alignements multiples plus informatifs que les ali. de 2 séq.

Les 2 grands types de MSA

- Un MSA peut être global ou local (comme un alignement de 2 séquences)

- Global : l'alignement 2 à 2 est étendu pour inclure 3 séq. ou plus
  Des protéines de $\neq$ organismes peuvent être conservées sur toute la longueur si elles assurent une fonction biologique importante

  Logiciels : CLUSTALW, MULTALIN, T-COFFEE, DIALIGN, . . .

- Local : recherche de domaines/régions conservés
  Les domaines fonctionnels de protéines peuvent être conservés tandis que le reste de la séquence diverge

  Logiciels : BLOCKS Web site, eMOTIF, GIBBS, HMMER, . . .

Sensibilité et sélectivité

- Notions utiles pour comparer des méthodes
- Par exemple : on prend une séquence membre d'une famille de protéine et on l'utilise comme requête, les algorithmes retourne une liste de *hits* et on coupe à un certain seuil

	Séq. membres famille	Séq. non membres
Séq. trouvé au-dessus du seuil	Vrais positifs	Faux positifs
Séq. trouvé en dessous du seuil	Faux négatifs	Vrai négatifs

- $\text{Sensibilité} = \frac{\text{Vrais positifs}}{\text{Vrais positifs}+\text{Faux négatifs}}$
- $\text{Sélectivité} = \frac{\text{Vrais positifs}}{\text{Vrais positifs}+\text{Faux positifs}}$

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
- Éditeurs de MSA
- Références

Comment quantifier la qualité d'un alignement multiple ?

- Généraliser une fonction de score pour un alignement de k séq.
- Le score d'un MSA est la somme des scores de ses colonnes (ici les colonnes sont de hauteur k)
- Les colonnes sont considérées indépendantes
- Fonction à k paramètres ?
  Si alphabet de taille $\alpha \Rightarrow \alpha^k$ colonnes possibles ($21^5 > 4$ millions)
  On ne peut pas associer un coût à chaque colonne ! ? !
- Mesure raisonnable S, quelles propriétés ?

Propriétés d'une méthode de score

1. Même score pour les colonnes contenant les mêmes caractères (indépendamment de l'ordre)

$$S(I,-,I,V) = S(V,I,I,-) = S(V,I,-,I) = S(V,-,I,I) = \cdots$$

2. Récompense les colonnes avec beaucoup de résidus identiques ou similaires

3. Pénalise les colonnes avec des résidus différents et des espaces

- Plusieurs méthodes de score : méthode SP, méthode basée sur la phylogénie (arbre ou étoile), contenu en information, . . .

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
					60

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24			

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

$p(C,C) = 9$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9		

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

$p(C,C) = 9$

$p(I,I) = 4$

$p(I,-) = -2$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9	16	

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9	16	40

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V	
2	N	N	N	I	V	
3	N	N	N	–	V	
4	N	N	C	I	V	
5	N	C	C	I	V	
	60	24	9	16	40	=

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V	
2	N	N	N	I	V	
3	N	N	N	–	V	
4	N	N	C	I	V	
5	N	C	C	I	V	
	60	24	9	16	40	= **149**

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

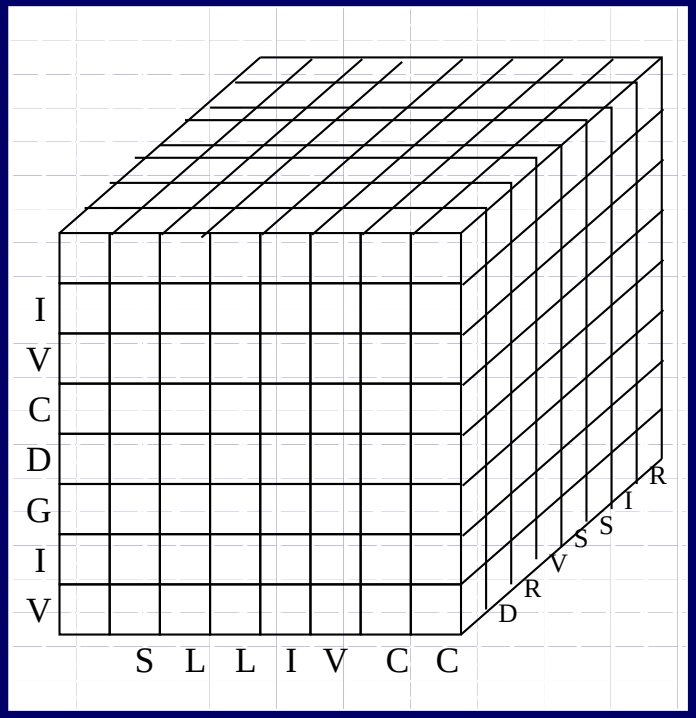
- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
- Éditeurs de MSA
- Références

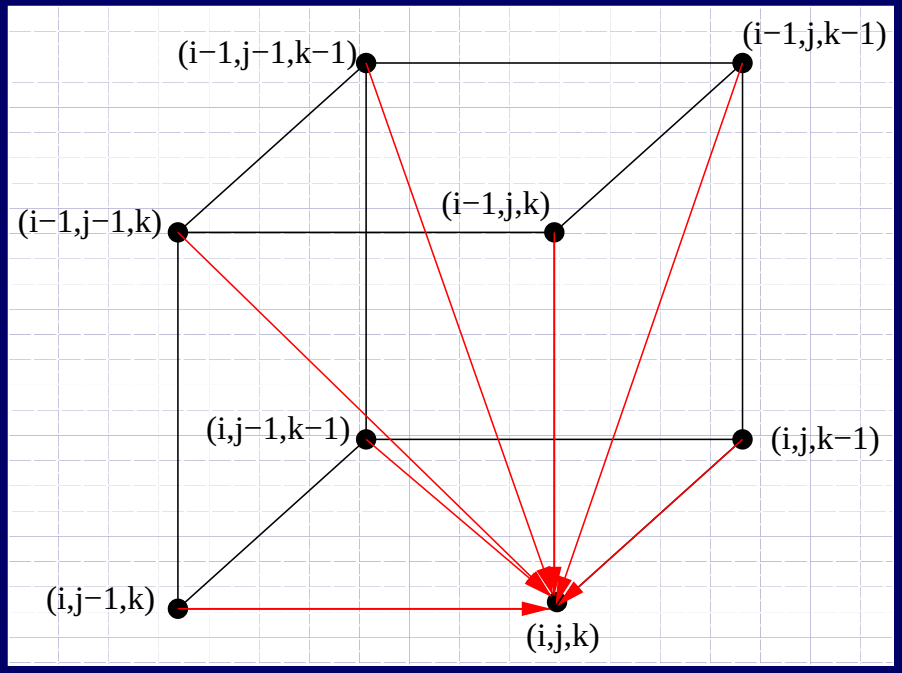
Méthode exacte

- Le principe de programmation dynamique est généralisable à k séquences de longueur n
- Matrice $\mathcal{A}$ de dimension n^k
- $\mathcal{A}(i_1, i_2, \ldots, i_k)$ contient le score d'alignement optimal entre les préfixes $s_1[1..i_1], s_2[1..i_2], \ldots, s_k[1..i_k]$
- Remplissage des $O(n^k)$ cases la table $\Rightarrow$ espace mémoire $O(n^k)$
- Chaque entrée dépend de $2^k - 1$ entrées déjà calculées
- Calculer le score SP requiert $O(k^2)$ car il y a $\frac{k(k-1)}{2}$ paires

Temps total d'exécution en $O(k^2 2^k n^k)$

Exemple avec 3 séquences





Exemple avec 3 séquences : la récurrence

- 3 séquences : r, s et t, méthode de score SP et g pénalité de gap
- $\mathcal{A}$ matrice de prog. dyn.

Initialisation : $\mathcal{A}(0,0,0) = 0$; $\mathcal{A}(i,0,0) = i \times 2g$;
$\mathcal{A}(0,j,0) = j \times 2g$; $\mathcal{A}(0,0,k) = k \times 2g$.

Remplissage :

$$\mathcal{A}(i,j,k) = \max \begin{cases} \mathcal{A}(i-1,j-1,k-1) & + & SP(r_i, s_j, t_k) \\ \mathcal{A}(i,j-1,k-1) & + & SP(-, s_j, t_k) \\ \mathcal{A}(i-1,j,k-1) & + & SP(r_i, -, t_k) \\ \mathcal{A}(i-1,j-1,k) & + & SP(r_i, s_j, -) \\ \mathcal{A}(i,j,k-1) & + & SP(-, -, t_k) \\ \mathcal{A}(i,j-1,k) & + & SP(-, s_j, -) \\ \mathcal{A}(i-1,j,k) & + & SP(r_i, -, -) \end{cases}$$

Méthode exacte : conclusion

- Méthode guère plus complexe pour k séquences que pour 2
- Facilement programmable
- Mais il est clair que le temps de calcul en $O(k^2 2^k n^k)$ devient prohibitif quand k augmente
- Illustration :
  - $\rightarrow$ 2 séquences de 100 a.a. $\rightarrow$ 1 sec.
  - $\rightarrow$ 3 séquences de 100 a.a. $\rightarrow$ 10 min
  - $\rightarrow$ 4 séquences de 100 a.a. $\rightarrow \sim$ 3 jours
  - $\rightarrow$ à partir de 9 séq. le tps de calcul dépasse l'âge de l'univers . . .

$\Rightarrow$ Mise au point d'algorithmes heuristiques performants et de bonne qualité (pb encore ouvert aujourd'hui)

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement progressif (ex : CLUSTALW)
  - → Alignement itératif (ex : DIALIGN)
  - → Alignement structurel & hybride (ex : STAMP, MULTALIN et T-COFFEE)
- MSA Local
- Éditeurs de MSA
- Références

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement progressif (ex : CLUSTALW)
  - → Alignement itératif (ex : DIALIGN)
  - → Alignement structurel & hybride (ex : STAMP, MULTALIN et T-COFFEE)
- MSA Local
- Éditeurs de MSA
- Références

Méthode progressive ou « alignement de profils »

- Difficile d'aligner simultanément k séquences $\Rightarrow$ on les aligne progressivement en commençant par les plus proches

$\rightarrow$ on aligne alors d'autres séquences avec ces premiers alignements

- Les alignement intermédiaires sont appelés « profils »

- À chaque étape on n'aligne que 2 éléments : séquence/séquence, séquence/profil ou profil/profil

- Utilise la prog. dyn. pour évaluer la similarité entre les séquences $\Rightarrow$ création d'un arbre évolutif qui sert de guide

- Propriété de cette méthode : toute création d'un trou est définitive « *once a gap, ever a gap* »

- Programme original CLUSTAL utilisé depuis 1988, régulièrement amélioré depuis [Higgins & Sharp, 88]
- CLUSTALW ([Thompson *et al*, 94]) est la version la plus récente, W pour *Weighting* ; les séquences et les paramètres sont pondérés
- Fonctionnalités : ajout d'une séq. ou d'un ali. à un ali. déjà fait, production d'un arbre phylogénétique, paramètre *slow/fast*, ...
- Méthode
  1. Construction de la matrice des $\frac{k(k-1)}{2}$ distances
  2. Logiciel de reconstruction phylogénétique $\Rightarrow$ arbre
  3. Alignement progressif suivant cet arbre
- Complexité pour k séquences de longueur n : $O(kn^2)$

CLUSTALW : le score

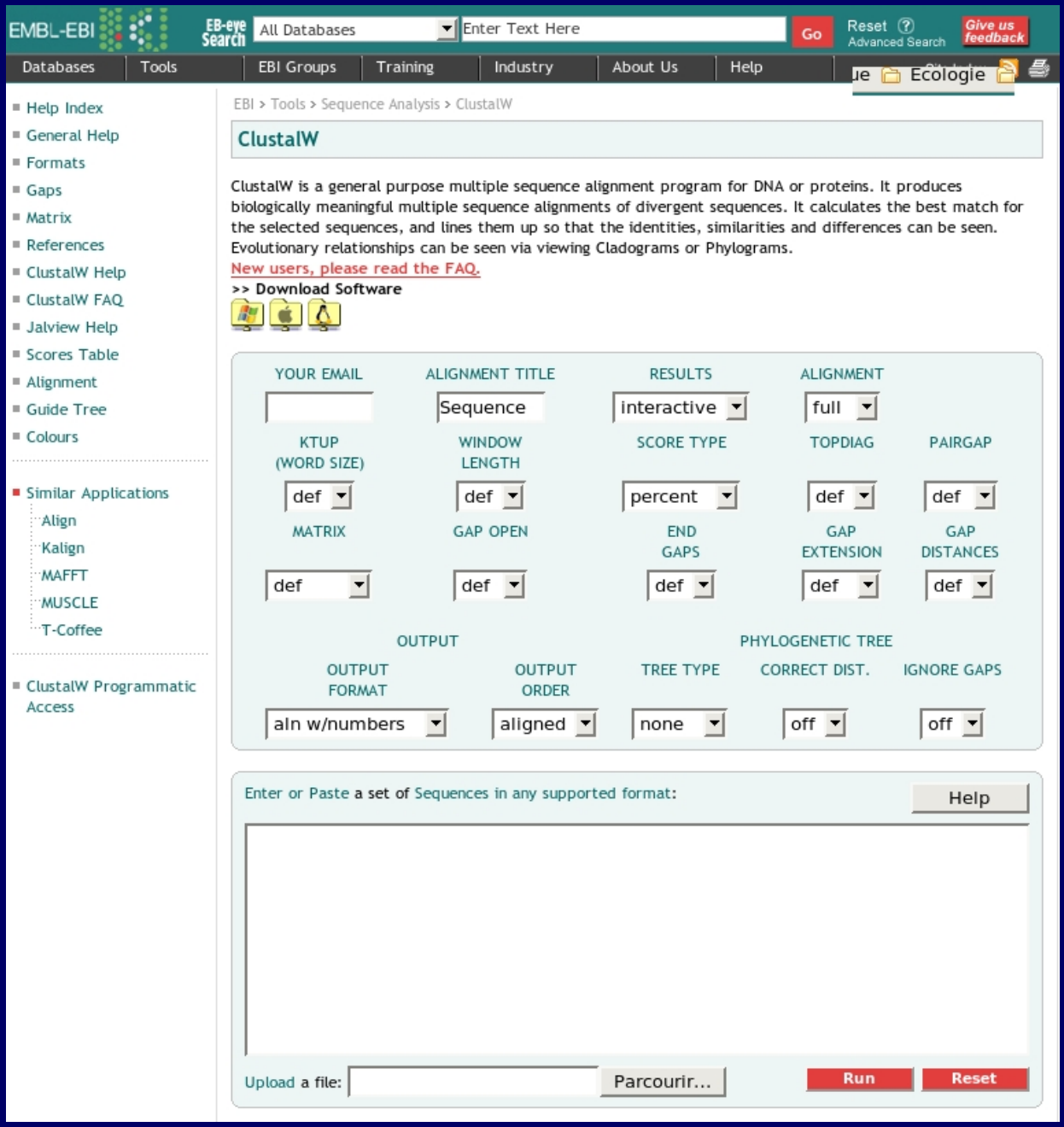
- Pondération des séquences :
  Une séquence similaire à d'autres dans le groupe a un poids faible, alors qu'une séquence moins proche des autres a un poids plus fort

  $\Rightarrow$ On diminue ainsi le poids des groupes de séquences similaires et privilégie les changements dans l'arbre évolutif

- Pénalité de gap :
  CLUSTALW pénalise les gaps de manière à les placer entre les domaines conservés (utilise une matrice spéciale et $\neq$ pénalités suivant les régions)

Méthode de score de CLUSTALW $\neq$ SP-score

CLUSTALW : Page d'accueil sur l'EBI



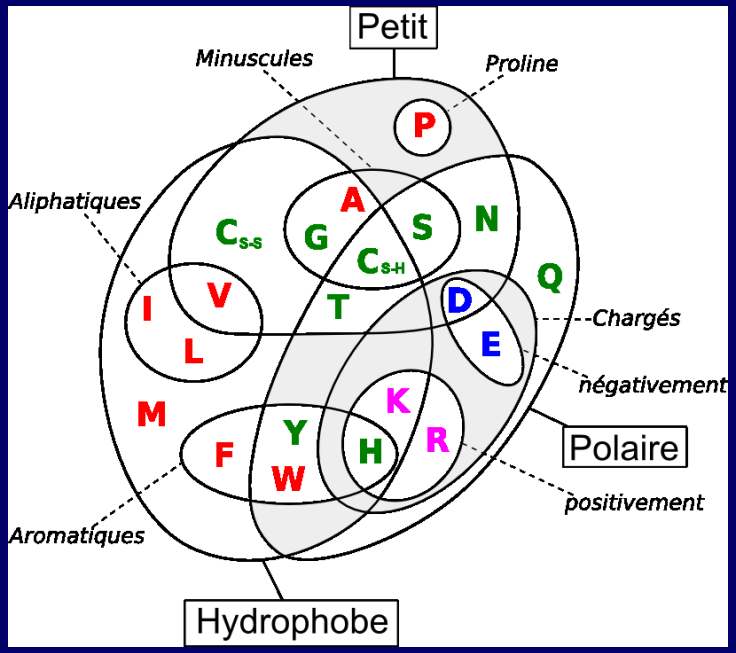
- CLUSTALW affiche par défaut les symboles suivants pour indiquer la degrés de conservation dans chaque colonne :

→ * caractère identique dans toute la colonne

→ : substitutions conservatives (suivant la table des couleurs)

→ . substitutions semi-conservatives

Code couleur de CLUSTALW



A.A.	Couleur	Description
AVFPMILW	Rouge	Petits (petits + hydrophobes (incl. aromatiques -Y))
DE	Bleu	Acides
RK	Magenta	Basiques - H
STYHCNGQ	Vert	Hydroxyl + sulfhydryl + amine + G
Autres	Gris	Unusual amino/imino acids . . .

Alignement étoile

- Un autre type d'alignement progressif : l'alignement étoile

- Principe : alignement 2 à 2 entre une séquence fixée (le centre de l'étoile) et toutes les autres séquences

- Méthode :
  1. Choisir la séquence centre s_c
  2. Alignement optimal entre toutes les s_i pour $i \neq c$ et s_c
  3. Agrégation des alignements avec le technique « *Once a gap, always a gap* », s_c est utilisée comme guide

- Procédure en $O(kn^2 + k^2l)$

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement progressif (ex : CLUSTALW)
  - → Alignement itératif (ex : DIALIGN)
  - → Alignement structurel & hybride (ex : STAMP, MULTALIN et T-COFFEE)
- MSA Local
- Éditeurs de MSA
- Références

- Le pb majeur avec l'alignement progressif est qu'une erreur faite au début de l'alignement ne peut être corrigée par la suite

- Les méthodes itératives tentent de pallier ce pb en réalignant des sous-groupes de séquences de manière répétée puis en alignant ces sous-groupes dans l'alignement global de toutes les séquences

- L'objectif est d'améliorer le score d'alignement global

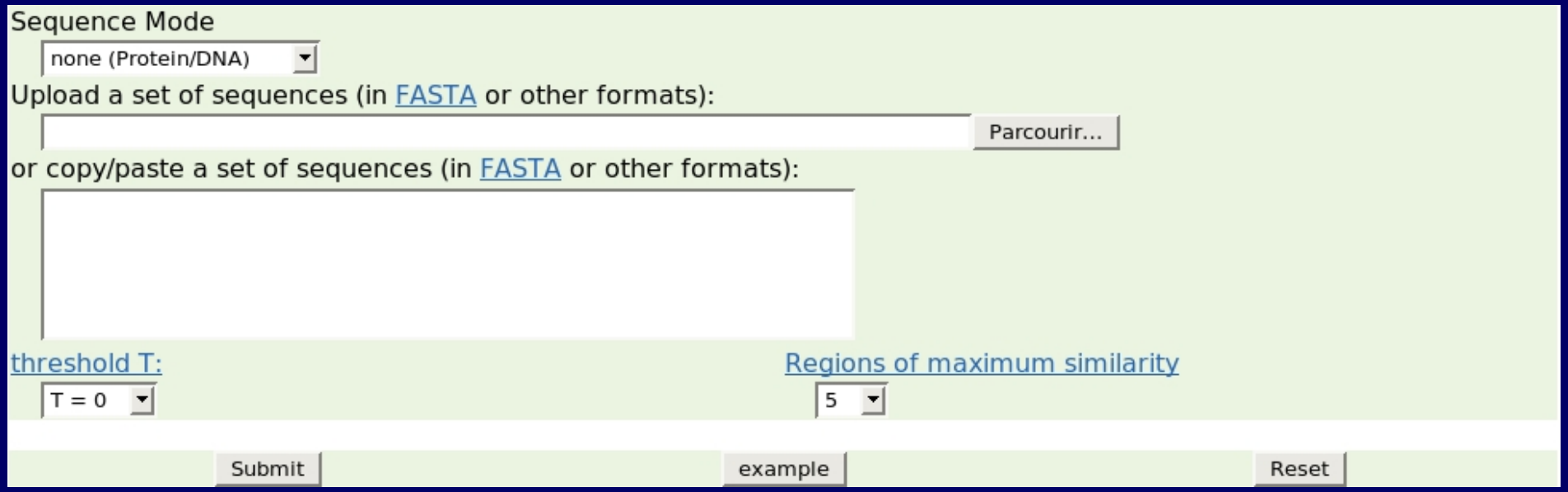
- La sélection de ces sous-groupes peut de faire sur la base d'un arbre phylogénétique (prédit de manière similaire aux méth. prog.), ou la séparation d'une à deux séquences du reste, ou de manière aléatoire

- DIALIGN 2.2.1 (oct. 2007) [Morgenstern, 2004]

- Spécialité : se base sur des similarités locales pour aligner des séquences très divergentes ou de longueur différentes

- Méthode :
  1. Repérer les régions alignées sans gaps dans les alignements 2 à 2 ($\sim$ diagonales continues dans un dotplot)
  2. Cherche un ensemble compatibles de diagonales pondérées pouvant produire un ali. et maximisant la somme des poids
  3. DIALIGN produit un alignement à partir de ces diagonales

- Disponible sur http://bibiserv.techfak.uni-bielefeld.de/dialign/



- Schéma de score dans DIALIGN : score de l'ali = somme des scores des diagonales qui le composent ⇒ pas de pénalités de gap
- Paramètres particuliers :
  - Seuil T → permet de ne considérer que les diagonales de score supérieur à T (nécessaire dans la V1)
  - Régions de similarité maximale → donne le nb max de caractères '*' utilisés pour représenter la similarité locale sous chaque col.

```
uniprot:MYH1    1898   SKFRRIQHEL EEAEERADIA ESQVNKLRVK SREVHTKIIS EE----
uniprot:MYH2    1900   AKFRKLQHEL EEAEERADIA ESQVNKLRVK SREVHTKVIS EE----
uniprot:MYH3    1899   SKFRKIQHEL EEAEERADIA ESQVNKLRAK SREIGKKAES EE----
uniprot:MYH3    1895   TKFRKAQHEL EEAEERADIA ESQVNKLRAK TRDFTSSRMV VHESEE
uniprot:MYH3     149   TKFRKAQHEL E--------- ---------- ---------- ------
uniprot:MYH3    1895   TKFRKAQHEL EEAEERADIA ESQVNKLRAK TRDFTSSRMV VHESEE
uniprot:MYH4    1898   AKFRKLQHEL EEAEERADIA ESQVNKLRVK SREVHTKVIS EE----
uniprot:MYH4    1897   SKFRKLQHEL EEAEERADIA ESQVNKLRVK SREVHTKVIS EE----
uniprot:MYH6    1896   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKQKM HDEE--
uniprot:MYH6    1896   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKQKM HDEE--
uniprot:MYH6    1896   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAK-KM HDEE--
uniprot:MYH6    1895   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKQKM HDEE--
uniprot:MYH7    1894   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGTKGLN EE----
uniprot:MYH7    1893   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKGLN EE----

                       ********** ********** ********** ********** ******
                       ********** ********** ********** *******
                       ********** ********** ********** ****
                       ********** ********** ********** ***
                       ********** *
```

- 3 formats de sortie possibles : FASTA, MSF et DIALIGN ci-dessus pour l'alignement multiple
- + un arbre au format PHYLIP
- DIALIGN peut aussi produire des alignements multiples locaux

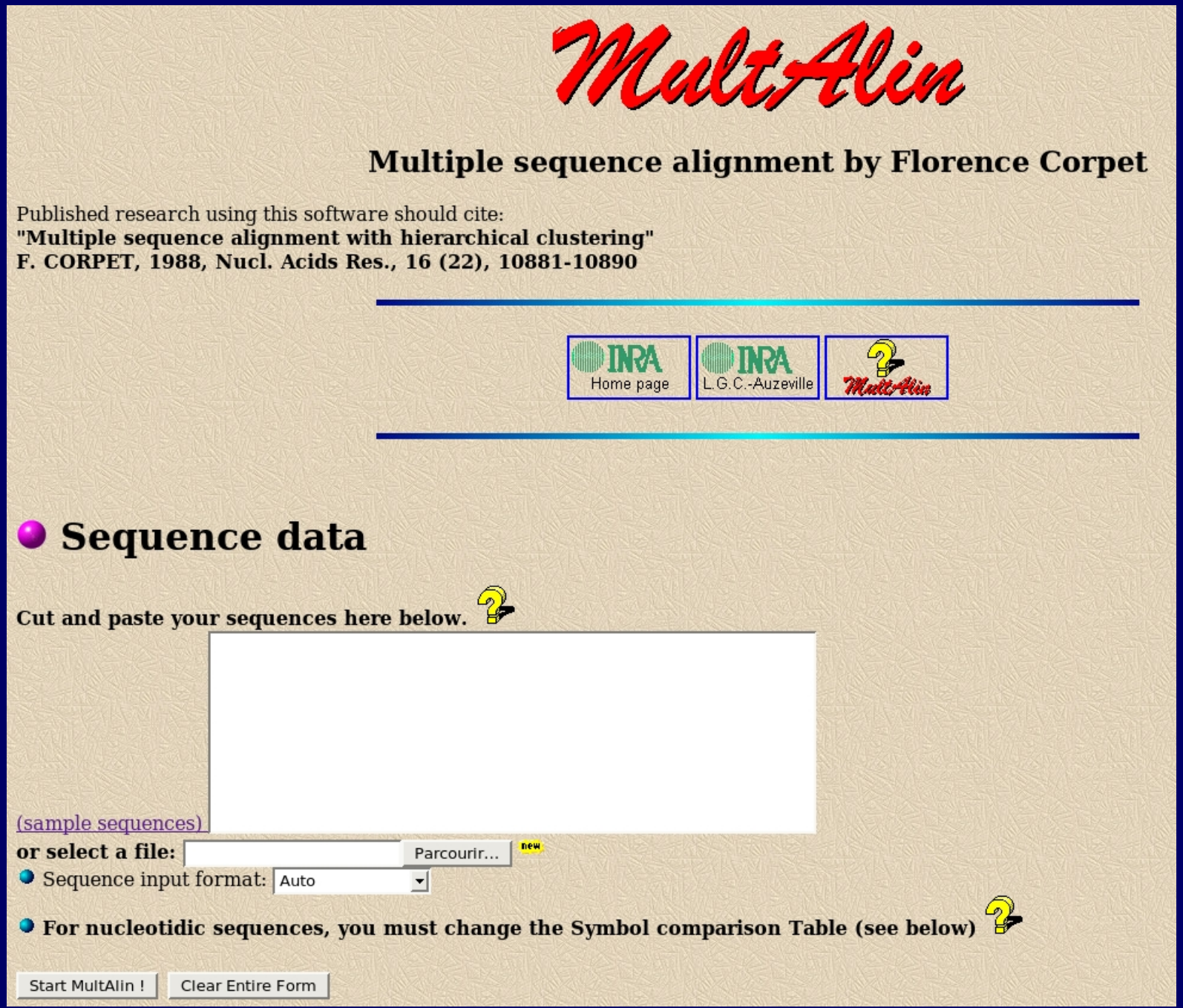
- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
  - → Alignement progressif (ex : CLUSTALW)
  - → Alignement itératif (ex : DIALIGN)
  - → Alignement structurel & hybride (ex : STAMP, MULTALIN et T-COFFEE)
- MSA Local
- Éditeurs de MSA
- Références

Alignement structurel

- Lorsqu'on dispose de structures 3D, on peut comparer des protéines en les superposant
- Mais il faut connaître quels sont les a.a. qui se correspondent
- C'est le but des méthodes d'alignement multiple structurel
  → Alignement structurel des séquences
  → Superposition 3D de leur structure
- Logiciels, pour 2 ou plusieurs protéines : SSAP, DALI, STAMP, ...
- Méthode de STAMP
  1. Alignement structurel de 2 séquences
  2. Les structures sont superposées selon cet alignement
  3. Calcul d'une matrice de scores de similarité structurelle
  4. Prog. dyn. sur cette matrice pour trouver meilleur score & ali.
  5. Étant donné l'alignement on recommence jusqu'à convergence

  Si plusieurs séquences, toutes les paires de structures sont comparées et un arbre est calculé, et on le suit comme pour un alignement progressif

- Programme de Florence Corpet de l'INRA Toulouse [Corpet, 88]
- Alignement multiple avec clustering hiérarchique
- Principe similaire à CLUSTAL mais pas d'évolution depuis 2000
- Méthode :
  1. Initialisation : tous les ali. 2 à 2 sont faits et on garde leur score
  2. Clustering hiérarchique des séquences à partir de ces scores
  3. Alignements selon cet arbre hiérarchique $\rightarrow$ ali. complet
  4. L'alignement est montré et on calcule son score SP (et donc le score de chaque paire de séquences alignées)
  5. Nouveau clustering hiérarchique avec ces nouveaux scores
  6. Si le nouveau clustering est différent du 1er, on peut recommencer (pas 3.) jusqu'à ce que ce soit stable

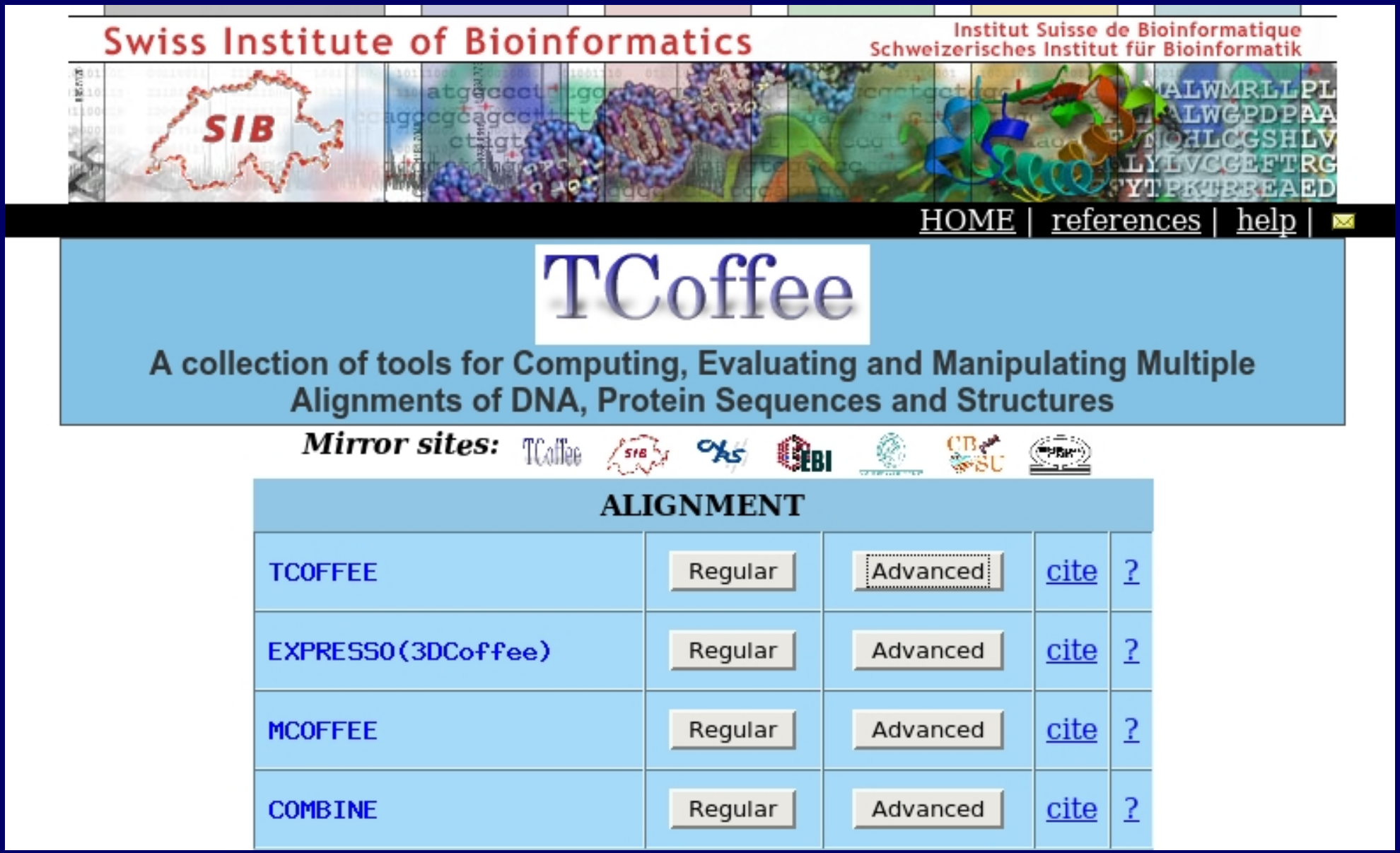

MultAlin
Multiple sequence alignment by Florence Corpet
Published research using this software should cite:
"Multiple sequence alignment with hierarchical clustering"
F. CORPET, 1988, Nucl. Acids Res., 16 (22), 10881-10890
INRA Home page
INRA L.G.C.-Auzeville
MultAlin
Sequence data
Cut and paste your sequences here below.
(sample sequences)
or select a file:
Parcourir...
new
Sequence input format:
Auto
For nucleotidic sequences, you must change the Symbol comparison Table (see below)
Start MultAlin !
Clear Entire Form

Résultats graphiques de MULTALIN

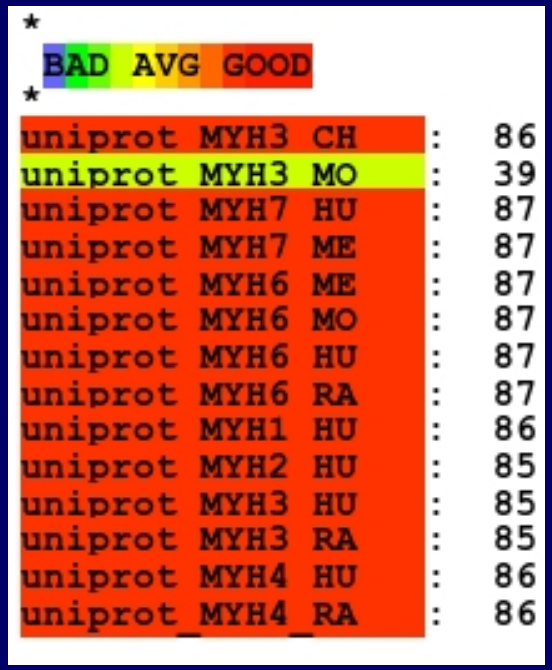
- Plusieurs formats d'entrée et de sortie disponibles
- Code couleur suivant le degrés de consensus de la colonne :
  - Rouge → fort consensus, lettre majuscule
  - Bleu → faible consensus, lettre minuscule
  - Noir → neutre, pas de caractère reporté
- Possibilité de modification des paramètres de la méthode et en particulier de ceux d'affichage

- T-COFFEE V5.05 (2007), [Notredame et al, 2000]
- Méthode :
  1. Construction d'une bibliothèque d'alignement 2 à 2, globaux et locaux, pour chaque paire de séquences du jeu d'entrée
  2. Phase de "scorage" pour donner un poids aux alignements puis aux caractères alignés (« *library extension* »)
  3. Cette bibliothèque est ensuite utilisée pour guider une phase d'alignement progressif pour trouver un MSA préservant la consistance des alignements 2 à 2
- Produit de meilleurs alignements mais encore trop lent pour de gros jeux de séquences
- Originalité : permet de combiner des résultats de CLUSTALW, DIALIGN et d'un alignement structural par exemple

T-COFFEE : Page d'accueil



T-COFFEE : Résultats graphiques



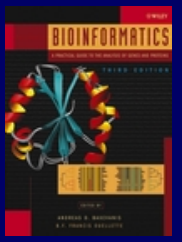
- Les parties oranges et rouges sont considérées comme bien alignées
- Plusieurs formats de sortie : .aln (CLUSTALW), FASTA, PHYLIP et T-COFFEE ci-dessus
- + 2 formats d'arbres

- Les MSA peuvent servir à faire des recherches plus sensibles dans les banques de données qu'en utilisant 1 seule séq. requête
- Rappel du fonctionnement de PSI-BLAST
  1. On cherche des séquences similaires à la requête
  2. Construction d'un profil ou PSSM
  3. (en boucle) On cherche avec ce profil et on y intègre les nouvelles séquences trouvées
  4. Fin qd l'ensemble de séq. est stables ou le nb max d'itérations est dépassée
- PSI-BLAST fournit des alignement multiples différents de ceux obtenus avec les outils classiques

→ Normalement les MSA sont plus long que les séquences qu'ils contiennent mais PSI-BLAST supprime les régions qui nécessiteraient d'introduire des gap dans la séq. requête

- Introduction
- Méthodes de score
- Alignement multiple exact
- MSA Global
- MSA Local
  - → Analyse de profils
  - → Analyse de blocs
  - → Méthodes statistiques
- Éditeurs de MSA
- Références

- Chap 13 dans “Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition” [Baxevanis & Ouellette,2005]



- “Bioinformatique Génomique et post-génomique” [Dardel & Képès, 2002]

- “Bioinformatics and Molecular Evolution” [Higgs & Attwood, 2005]

- Illustrations :

  → Transp. 22 : Diagramme de Venn des propriétés des acides aminés. Wikipedia, entrée ’Acide aminé’ (http://fr.wikipedia.org/wiki/Acides_aminés)